知能と情報（日本知能情報ファジィ学会誌）Vol.21, No.4, pp.519-526（2009）

解説

自然言語テキスト意味概念の共通的記述による次世代Web基盤†

石塚 満*1・内田 裕士*2・横井 俊夫*3

1．まえがき

情報は今後も社会や個人活動の様式を左右する重要な要素であり，それらを変革していく原動力である．インターネット，Webにより情報及び知識の流通，グローバルな共有，更には集合知のような共創の形態は大きく変化した．流通，蓄積される情報量は今後とも増大することは確実であり，その利活用法が個人，組織，社会，そして国の活力，創造力，競争力に反映することになる．大規模，大量の情報を扱うとなると，コンピュータの力を借りることが不可欠となる．キーワード検索を基本とするWeb検索エンジンは情報アクセスに革命的といえる変化をもたらし，またWeb等におけるデータマイニング，特に統計的データの取得と利用も日常的になってきた．これらは今後とも必要な機能であるが，情報を表層的にとらえて処理するレベルである．人間は情報の表す意味も把握し処理するのだが，人間レベルの意味理解はすぐには難しいとしても，意味レベルに幾分踏み込んで大量情報の処理，操作を可能とすることは次世代Web基盤として重要な方向である．

Webの生みの親であるTim Berners-Leeの提唱により1999年から進められているSemantic Webも，このような方向でWebコンテンツをコンピュータ可読な形式にすることを目指したものである．その記述の対象はメタデータであり，W3C(World Wide Web Consortium)において3つ組構造データの記述形式であるRDF(Resource Description Framework)(注1)，使用する語彙及びオントロジー定義の形式としてRDFS(RDF Schema)，OWL(Ontology Web Language)が標準化されている．Web上の意味計算(Semantic Computing)基盤への第一歩として期待されてはいるのだが，RDFの利用は広がってきているものの，オントロジーも含むSemantic Web全体としての広がりは必ずしも十分でなく，課題も見えてきている．

幾つか挙げると，Semantic Webの基本語彙セット(オントロジー)は領域毎に構築されるので，領域間に亘る統合が困難であり，広がりが達成されない．領域間でオントロジーを写像する研究も行われているものの，成功するのは僅かな部分に限られている．オントロジーを記述するOWLは記述論理(Description Logic)に基づいて設計されているが，一般ユーザが理解し使用するには，敷居が高くなり過ぎている感がある．Semantic Webの名称自体は魅力的であるが，対象とするのはメタデータであり，Data Web(コンピュータ可読なデータベース化されたWeb)といった呼称の方がその機能を適切に表している(T. Berners-Leeも最近はそのように述べている)．

本稿で紹介するのは，CDL(Concept Description Language)／CWL(Common Web Language)に基づく意味計算(Semantic Computing)基盤についてであり，次世代Web基盤へ貢献すべく努力している日本発の技術である．Semantic Webがメタデータの共通的記述を行うのに対し，CDL／CWLは自然言語テキストが表す概念をコンピュータが把握できる形式で共通的に記述する言語である．(CDLは画像，映像なども含むコンテンツの意味記述への配慮もなされているが[横井06]，それらの意味も多くの場合に自然言語を介して記述されるので，ここでは自然言語テキストを記述対象とする．)即ち，メタデータには限定せず，Webで情報／知識伝達の主要なソースデータであるテキストが表す意味概念を記述対象にしている．更に，Semantic Webのオントロジー構築が領域毎で問題があると述べたが，CDL／CWLの語彙はユニバーサルに

† A Next-generation Web Foundation based on the Common Description of Concepts Expressed in Natural Language Texts
Mitsuru ISHIZUKA, Hiroshi UCHIDA and Toshio YOKOI

*1 東京大学 情報理工学系研究科 創造情報学専攻／電子情報学専攻
School of Information Science and Technology, the Univ. of Tokyo
*2 UNDL財団
UNDL Foundation
*3 (財)日本特許情報機構(JAPIO)特許情報研究所
Japan Patent Information Organization(JAPIO)

(注1) RDFはSemantic Web以前にR. V. GuhaとT. Bray(当時Netscape)により開発されたものである．

定義され，体系的に構成されている．これら語彙を結びつけて複雑な概念，そして文の表す意味概念を表すのに用いる関係子も，特定の言語に依存することなくユニバーサルに定義されている．これによって，言語の壁を越える情報／知識の交流を可能にする．

英語が支配的であるグローバルなWebの情報流通・共有において，言語の壁を克服し多言語による交流を可能にすることは今後の重要な課題である．しかし，英語を主とする国ではこの課題に対して関心が薄く，解決の技術は生まれてこない．日本のような非英語，非印欧系言語の国が主導して，国際貢献を果すべき領域と言える．

言語の壁の克服というと第一に機械翻訳を思い浮かべよう．機械翻訳あるいはコンピュータによる言語意味理解は着実に進歩しているものの，人間の能力との間のギャップは依然として解消されず，この解消は(永遠に)不可能ではないかとも言われている．CDL／CWLでは，自然言語テキストの表す概念を十分近似的に表現する適切なレベルを設定し，機械翻訳のようにテキストからの全自動変換を目指すのではなく，人手による(負荷の軽い形態での)援助を含む形でテキストからCDL／CWLへ半自動変換する．

意味表現といっても深層から表層まで幾つかのレベルがあるが，深層レベルはどのような要素によって表すかについてのコンセンサスがなく，まだ共通化は困難である．CDL／CWLで扱うのはテキストの表層表現から僅かに意味に踏み込んだレベル(概念表現レベルと称する)であり，具体的には各語彙要素の意味役割(semantic role)を明示化したレベルと位置付けられる．このレベルは長年の自然言語処理・理解研究により一定程度の共通的理解ができてきており，今日，標準化を図る意義は大きいと言える．

自然言語テキストの意味表現，あるいは知識表現として利用に関する他の試みについては[石塚06]に紹介されているので，参照願いたい．

2．CDL／CWL開発の経緯

CDL／CWLに先行してUNL(Universal Networking Language)の開発があり，その基になったのは多言語機械翻訳技術の中間言語(ピボット言語)であった．図1はこのような経緯を示している．

UNLは1996年にスタートした国連大学のUNLプログラムのもとで開発された[Uchida05，UNDL]．主な目的はWeb上での言語の壁を越える情報の交流を可能にすることであり，基になっているのは多言語機械翻訳における中間言語[Uchida 90]である．

CDLは単にテキストだけでなく，今後のWebにおいて他も含む広いメディア一般が表す概念を記述する汎用で基本的な枠組み，及びWebを中心とした今後の意味計算基盤として設計された[Yokoi05，横井06]．CDL.nl(CDL natural language version)[CDL.nl 06]はUNLの成果をCDLの枠組みの上に移し再設計したものであり，自然言語テキストの意味する概念を汎用的に記述する言語である．

CDL.nlは次世代Web基盤として世界に認知されかつ世界標準として貢献する必要性から，2006年からW3C(World Wide Web Consortium)で標準化活動を開始するのに際し，より分かりよく受け入れられやすい名称にするために名付けたのがCommon Web Language(CWL)である[CWL08]．実はCWLはアンブレラ仕様であり，CWL.unl(UNL記述に対応)，CWL.cdl(CDL記述に対応)，CWL.rdf(RDF表現を採用)を包含している．UNLを利用する活動は世界数カ国で行われており，UNLとその関連機能の意義も高いのでCWLの核に含めている．CWL.rdfはRDFでCDL.nl記述を表現したものであり，記述自体は長いものになっ

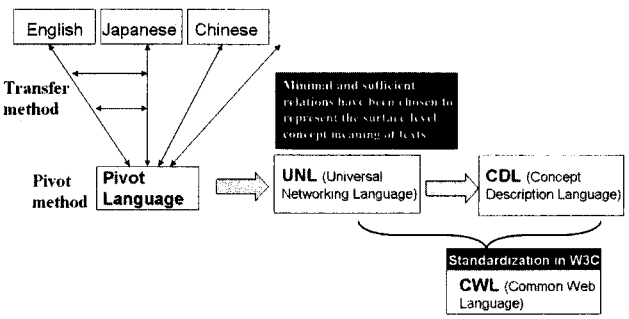


図1　CDL／CWL開発の経緯

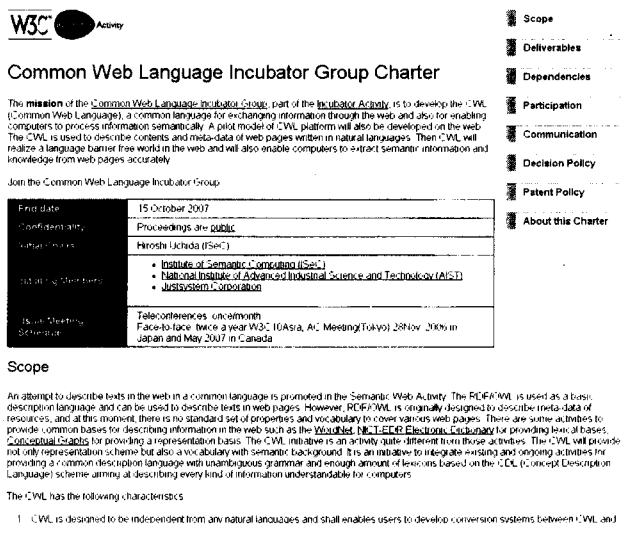


図2　W3Cインキュベータグループの憲章(charter)部分

てしまうが，RDF関連ツール類の利用が可能になるメリットがある．

W3Cでの標準化はまだ準備段階のインキュベータグループ活動になっており，2006年10月から2008年5月までの第1期でCWLの仕様を定めて公開した[CWL 08]．その後の第2期で関連ツール類を含めて評価を行っている．標準化に際してはSemantic Web標準との違いを強調するのでなく，メタデータも記述対象の一部と考え，むしろ整合をとるようにしている．図2はこのインキュベータグループ憲章(charter)のWebページを示している．

3．UNL

UNLはハイパーノード(ノード中にネットワークを含むことが出来る)を含む意味ネットワークの形式で，自然言語テキストが表す意味概念を曖昧性なく表す．ノードは概念を表し，アークは概念間の関係を表す．概念には注釈が付けられる．

UNLの語彙を構成するのがUW(Universal Word)である．UWは概念を表す"単語"であり，他のUWと関係子によってリンクされ複合概念，そして文の意味を表現する．関係子は文を構成する各々の単語(UW)の主に意味役割を指定することになる．作者が意図した主観的な意味は属性で表される．このモダリティを含む主観性(subjectivity)を属性として記述することで，UNLは文のニュアンスも含めて表現可能である．

UWは世界中の人々が認識できる概念を正確に表現し，かつそのような概念が誰でも同じ記号でもって定義できるようになっている．意味素には便宜的に英語の単語表現を借用して用いるが(人工的記号にすると人が見て理解出来なくなってしまう)，意味素の表現する概念の範囲を限定し，表現したい概念の範囲を正確に表現するために，UNLの関係子に基づく他のUWからの束縛(意味的包含関係と意味的共起関係)によって規定している．これにより多義性をもつ単語も，多義を分解した一義の語義(word sense)として定義していることが大きな特長である．

例えば"spring"は語義毎に以下のように規定されている．

[名詞概念]

1) spring(icl>tool)：バネ，スプリング
2) spring(icl>season)：春
3) spring(icl>fountain)：泉，湧水

[動詞概念]

4) spring(agt>thing，obj>wood)：反り返らせる
5) spring(agt>thing，obj>mine)：爆発させる
6) spring(agt>thing，obj>person，src>prison)：出獄させる，釈放させる
7) spring(agt>thing，gol>place)：跳ぶ，跳ねる
8) spring(agt>thing，gol>thing)：飛び越す
9) spring(obj>liquid)：湧き出る

簡単に記号の説明をすると，"icl"は(includedで)subclass関係を表し，"agt"，"obj"，"src"，"gol"はそれぞれ動作主，対象，始状態，終状態の関係子を表している．

このようにしてUWは語義単位に意味や役割は曖昧性なく知識ベース(UNLKB)に定義されている．現在，UWの語彙数は約20万語である．新しい語は既存の語彙との関係を明確にして付け加える．各言語の語彙は上記右側に対応日本語単語が記してあるように，このUWに対応させて対訳辞書とする．(UNLから自然言語テキストを生成する際にUWの訳語が複数ある場合は，共起辞書を参照して訳語選択を行う．)UW語彙は言語の違いを越えてユニバーサルに使用されるものであり，Semantic Webに於けるような異なるオントロジー間の変換といった問題は生じない．このような体系的語彙オントロジー(注2)構成法は，この観点からも今後重要になると思われる．

UNLのモダリティや関係子の種類についてはCDL.nlと共通なので後述する．各言語からUNLへの変換支援システム(Enconverter)，逆にUNLから各国語テキストを生成するシステム(Deconverter)は各国語毎に開発が必要であり，英語，日本語，スペイン語，中国語，仏語，アラビア語など十数カ国の言語について各国で協調し開発されている．継続して性能向上が必要であり，現在は部分的にCDL／CWLと歩調を合わせた開発が行われるようになっている．

UNLについては他にも記すべきことがあるが，そのWebページ(http://www.undl.org/)を参照していただくとして，CDL.nlと共通部分は以下の節で記す．

4．CDL

CDL自体は自然言語テキストだけでなく，他も含む広いメディア一般が表す概念を表現する汎用で基本的な枠組みとして設計されている[Yokoi 05，横井06]．簡素な3つの組表現(<実体1，関係，実体2>，<主語，述語，目的語>あるいは<実体，属性，属性値>などを表す)を基礎とし，ネットワーク表現だと実体を表すノードと関係を表すアークから成り，ノードは(更には一般にはアークも)ネットワークを含むことが

(注2) オントロジーは幾分異なる意味をもって使われることがあるが，ここでは相互に関係が定義された基本語彙セットである．

出来る(複合実体となる)のでハイパーノードとなる．関係構造とハイパーノードに対応する入れ子構造の表現を基本にしている．実体と関係には詳細特性記述のために任意個の属性ー属性値(attribute-attribute value)を付加できる(関係に準じるのものであり，もし属性値が複合実体となるなら関係とすることができる)．

CDL.nl [CDL.nl] はUNLの成果をCDLの枠組みの上に再設計した，自然言語テキストの意味する概念を汎用的に表現する概念記述言語である．UNLの成果を活用し，自然言語表現をカバーする実用レベルの基本語彙オントロジーと関係子を備えていることも大きな特徴である．Semantic Webにおけるメタデータの記述に留まらず，Webの主要情報ソースである自然言語テキストを対象にして共通的な意味概念記述を可能とする．そして次世代Web基盤としての国際標準化を視野に入れ，十分な汎用性，国際性を持つことを前提にして設計されている．図3はCDL.nl表現と対応するネットワーク表現を例示している．ここではthat以下の構文がハイパーノードになっている．

自然言語の意味に関しては表層から深層まで幾つかの階層があるが，CDL.nlで記述対象とする意味概念は，表層表現から僅かに意味に踏み込んだ一般性がある概念レベルとしている．意味表現というと深層意味ととらえて，正しく理解できないだろう，共通的な記述法は無理ではないかといった質問をよく受けるが，深層意味を記述の対象にしている訳ではない．幾分具体的に表現対象を規定すると，次のようである．

1) 辞書的多義を分解した概念を要素実体概念，事物概念とする(UNLのUWに対応)．
2) 事物概念を要素実体概念として，述語成分と格成分及び述語修飾成分によって，複合実体概念，単事象概念を構成する．

- **<John reported to Alice that he bought a computer yesterday.>**
- {#A01 Event tmp='past';
  {#B01 Event tmp='past';
  {#b01 buy;} {#b02 computer ral='def';} {#b03 yesterday;}
  [#b01 agt John] [#b01 obj #b02] [#b01 tim #b03] }
  {#John John;} {#Alice Alice;} {#a01 report;}
  [#a01 agt #John] [#a01 gol #Alice] [#a01 obj #B01] }

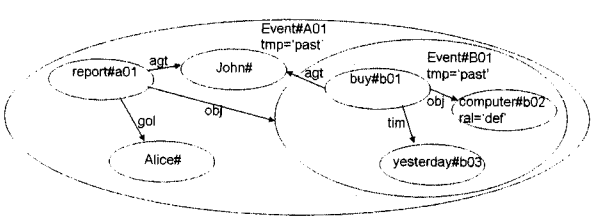


図3 テキスト(上段)，そのCDL.nl記述(中段)とネットワーク表記(下段)の例

3) 単事象(単文)概念は命題部分とモダリティ的部分に分ける．
4) 単事象(単文)概念を要素実体概念として節の修飾関係や主従関係によって複合実体概念，複事象(複文)概念を構成する．
5) 複事象(複文)概念を要素実体概念として，それらの論理的・時間的・因果的関係や相互参照関係によって複合実体概念，状況(文章・談話)を構成する．

文は命題部分とモダリティ的部分に分けて記述するが，命題部分は“実体ー関係”で表し，モダリティ的部分(アスペクト，テンス，ポラリティ，モダリティ)は“属性ー属性値”で表す．図3ではtmp=“past”(事象のテンスが過去を表す)が属性となっている．命題部分は，述語成分，述語修飾成分，複数の格成分という実体が格関係あるいは意味役割で結ばれた構造となる．それぞれの成分が複合語，句，節で構成されている場合には複合実体となる．文の実体内には格関係，意味役割に加えて，内部での参照構造を表す関係も記述することになる．

文章は複数文から構成されるが，文章の複合実体は複数の文の複合実体をノードとして含み，それらのノード間は文間の接続構造を表す接続関係か，指示表現や代用関係等を表す参照関係で結ばれる．(しかしCDL.nlの文間関係(談話関係)記述子についてはまだ完全に仕様化されておらず，今後の課題になっている．)

基本語彙(記述に必要な関係子語彙セットを含む)はUNLの成果を受け継ぐ形で構成されているが，これが充実していることが大きな特長であることから，ここで全体を示すことは出来ないものの，その一端を示すことにする．

CDL.nlの語彙を分ける最上位オントロジーは，

Entity，Relation，Attribute

である．Entityの下位オントロジーにはThing(事物)とCompositeEntity(複合実体)がくる．Thingの下位に以下のカテゴリーが置かれる．

- NominalThing(抽象物，具体物，機能，有意志体，場所，代名詞を含む)
- NominalModifier(限定詞と形容詞)
- VerbalThing(Do，Occur，Be関係の動詞，述語)
- VerbalModifier(数量と副詞)
- ValueOfAttribute(属性値)

これらカテゴリーの下に現在約20万語がUNLのUWの形で登録されている．

Relation(関係)については，自然言語で表される概念をどのような観点からとらえるかを示す意味で重要

表1 CDL.nlの関係子語彙

ElementalRelation 要素関係
- **CaseRelation** 格関係 (=IntraEventRelation 事象内関係)
  - [QuasiAgent 準主体] Agt(agent：動作主), Aoj(thing with attribute：属性主), Coag(co-agent：並行動作主), Cao(co-thing with attribute：並行属性主), Ptn(partner：相手)
  - [QuasiObject 準客体] Ben(beneficiary：受益者), Cob(affected co-thing：並行対象), Obj(affected thing：対象), Opl(affected place：場所対象)
  - [QuasiInstrument 準方法] Ins(instrument：道具), Met(method or means：方法), Man(manner：仕方)
  - [QuasiState 準状態] Gol(goal, final state：終状態), Src(source, initial state：始状態), Via(intermediate place or state：経由)
  - [QuasiPlace 準場所] Plc(place：場所), Plf(initial place：起点), Plt(final place：終点), Scn(scene：場面)
  - [QuasiTime 準期間] Dur(duration：期間), Tim(time：時間), Tmf(initial time：始時間), Tmt(final time：終時間)
  - [QuasiBasis 準基準] Bas(basis for expressing a standard：基準)
- **InterThingOrInterEventRelation** 間事物・間事象関係
  - And(conjunction：連結), Con(condition：条件), Coo(co-occurrence：同起), Fmt(range/from-to：範囲), Frm(origin：起点・起源), Or(disjunction, alternative：選択), Pur(purpose or objective：目的), Rsn(reason：理由), Seq(sequence：先行), To(destination：目的地)
- **Quantification&ModificationRelation** 限定・修飾関係
  - Cnt(content, namely：内容), Man(manner：仕方), Mod(modification：限定), Nam(name：名前), Per(proportion, rate of distribution：単位), Pof(part-of：部分), Qua(quantity：量),

CompositeRelation 複合関係
- Pos(possessor：所有者)

表2 CDL.nlの属性と属性値

- **Time (with respect to writer)** (時制)
  - past present future
- **Aspect (view on aspect of event)** (事象の相)
  - begin complete continue custom
  - end experience progress repeat state
- **View of emphasis, focus and topic** (強調，フォーカス，話題)
  - contrast emphasis entry qfocus theme
  - title topic
- **Attitudes (modality)** (態度)
  - affirmative confirmation exclamation
  - imperative interrogative invitation politeness
  - respect vocative
- **Feeling and judgements** (感情と判断)
  - ability get-benefit give-benefit conclusion
  - consequence sufficient grant grant-not
  - although discontented expectation wish
  - insistence intention want will need
  - obligation obligation-not should unavoidable
  - certain inevitable may possible probable
  - rare regret unreal admire blame
  - contempt regret surprised troublesome
- **View of reference** (参照ビュー)
  - generic def indef not ordinal
- **Logicality** (論理的性質)
  - transitive symmetric identifiable disjoint
- **Modifying attribute on aspect** (相の修正)
  - just soon yet not
- **Attribute for convention** (記号表現規約など)
  - passive pl angle_bracket brace
  - double_parenthesis double_quote parenthesis
  - single_quote square_bracket

であるので，表1にその関係子語彙を提示する．

意味役割ラベルの例は他でも見られ，例えば英語についてはPropBank [Palmer 05]で用いられている意味役割ラベルがある．しかしこれは英語動詞毎に役割を変えて規定する部分があり，言語独立のユニバーサル性はない．また，意味役割だけでは文の要素である実体概念間の関係を記述することは出来ず，論理関係，概念関係(上位や具体化など)，接続関係，参照関係も必要になり，表1のようにCDLはこれらの関係子も含んでいる．CDLの関係子が本当に必要最小限かを証明することは難しいが，多言語機械翻訳の中間言語，UNLによる記述を通じて実践的に十分であると実証されてきている．

ノード(実体概念)に付加される属性及び属性値は表2のように与えることができ，文の命題部分だけではなく，モダリティ部分も十分な精度で表現できるようになっている．他にこのように充実，整備された表現機能を有するシステムは存在しない．

以上で分かるように，CDL.nlは記述形式を定めているだけでなく，様々な自然言語テキスト表層表現の意味を，高い近似度の概念レベルで表現するのに必要な関係子語彙と属性を規定して提供している．これによってコンピュータは，曖昧性なく意味概念を把握できることになる．より深い意味理解や推論機能等は外部プログラムに委ねられる．すなわち，自然言語テキストは一旦CDL.nlの概念記述に変換され，これを共通的基盤として幅広い意味計算(セマンティックコンピューティング)の展開が可能になる．

CDL.nl自体は個別の言語に依存せず，汎用的に自然言語の意味概念表現に使えるものであるが，共通である関係子語彙(機能語)を除く語彙オントロジーを各国語対応に用意することにより，各国語対応バージョンを作成できる枠組みとなっている(日本語版のCDL.jp，中国語版のCDL.chなど)．各国語の語彙は基本とするUNLのUW(Universal Word)に対応付けることにより，CDL.nlは言語の壁を超える情報／知の交流の基盤ともなる．

5．テキストからの半自動変換

CDL.nlの現状での一つの課題は，その記述生成法である．自然言語テキストからの全自動変換は理想ではあるが，正しい機械翻訳と同様に困難であり，何らかの人手による援助が必要となる．この人の労力は出来るだけ負担が軽いものである必要がある．現在の開発状況を紹介する．

第一は手動による各単語の語義選択による方法である．この変換法は図4に示すように，UNLシステム

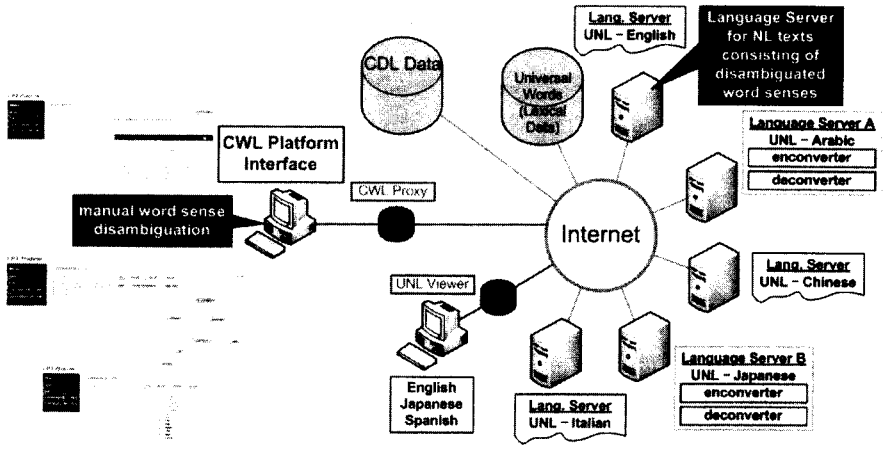


図4 人手語義選択によるテキストからCDL.nlへの半自動変換システムの構成

の機能を利用して実現されている．UNLシステム内で英文テキストからUNL記述に変換するのに基本的にルールベースの翻訳機構を用いているが，このルールセットは通常用途のものに加えて，単語語義曖昧性を解消し一義になった語義を対象とする翻訳ルールセットも備えている．後者のルールセットを用いることが出来れば，テキストからCDL.nlへの変換の精度を大幅に高めることができる．そこで，我々は多義で語義曖昧性を有する英語単語に対して，人手による語義選択，確定を行うユーザインタフェースを開発した．図3左の3枚のスクリーンショットはこのインタフェース画面のものであり，上から順に人手による語義選択画面，変換後のCDL.nl(或いはそれを包含するCWL.unl，CWL.cdl，CDL.rdf)のグラフ表示画面，CDL.nl記述表示画面を示している．全ての多義を持つ単語に対して人手による語義選択を要請するのは煩雑過ぎ，前後の文脈から語義を決定できる場合も多いので(例えば名詞のbankはお金に関する文脈中では銀行に相当する語義(UWの表現でbank(inl>organization))になる)，自動的に単語語義曖昧性解消の可能な部分を増やすようにしている．日本語かな漢字入力変換でも完全自動化は困難だが，今日は誰でもが"かな"から変換の漢字候補中から人手選択を行うことにより，人に無理のない形でかな漢字入力変換が実現されている．このアプローチの目指すところは，かな漢字入力変換のように軽負担で人手が関与する形態でのテキストからのCDL.nlへの変換である．

係り受け解析とまた十分ではない意味役割解析のツールを利用して，文中のエンティティ間に存在するCDL.nl関係を自動抽出する研究も行っている[Yan 08]．しかしこれは十分な性能になるまでにはまだ研究が必要である．特にCDL.nlで記述されたデータ(コーパス)がまだ十分な量でないことから，出現頻度が低いCDL.nl関係に関して機械学習(Support Vector Machine等の手法)を使って識別器を得るのが難しい．この問題に関しては，記述化データの増加と，CDL.nl記述化されていないデータも学習に利用する半教師付き学習(semi-supervised learning)の方法を導入して，状況の改善を図っている．

別のアプローチは，通常の変換ルールセットを用いてCDL.nlに変換し，間違い部分を後編集によって修正する方法である．これは機械翻訳の後編集に近いが，英語テキストからこの方法でCDL.nl記述に変換したデータは，グラフ表示と共に，CDL.nlから逆変換して英文テキストとして表示することで，原文テキストとの比較が容易であり，修正点を見出しやすい．機械翻訳の場合は，例えば英文と日本文といった異なる言語間で比較し，修正点を見出して後編集となる．

6．意味的検索

CDL.nl記述したデータはコンピュータも意味が把握出来ることにより，コンピュータも動員する広範な意味計算の基盤になるのだが，最も直接的な効用は意味的検索ということになる．現在のキーワード照合を基本とする検索を超えて，まず実体(entities)間の関係を指定した検索が可能になる．これは問い合わせもグラフで表すことにより，一種のグラフ照合によって実現できることになるが，まず完全な照合がとれるパターンを検索し，それが得られないときは単語を下位語，上位語，関係語に置き換えて探索することにより，意味的類似性に基づく検索が実現される．図5はこのような考えで実現したCDL.nlデータの検索システムのプロトタイプを示している．

ここでの問い合わせ言語はSQL-likeなCDQL(Concept

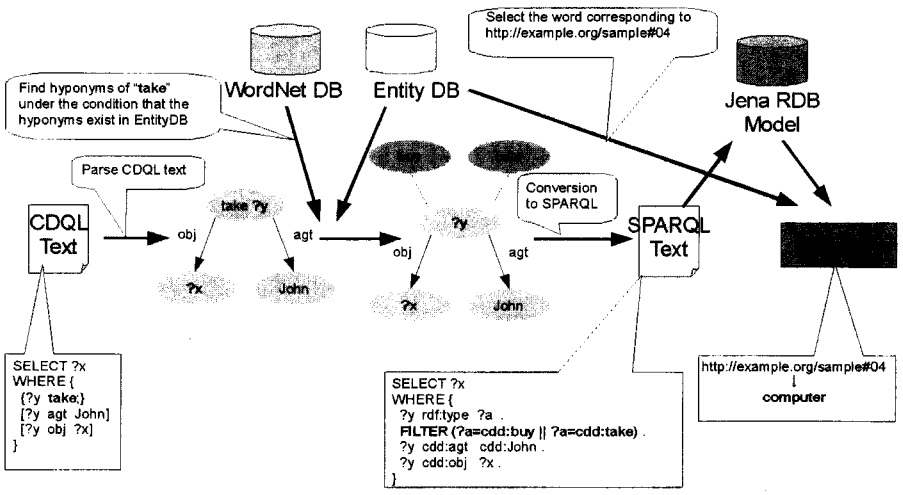


図5 語の意味的類似性に基づく検索の様子

Description Query Language)であり，処理系の内部ではRDFの問い合わせ言語として標準化されているSPARQL [SPARQL]を用いている．SPARQLはRDFデータを関係データベース(Jena RDB)で管理しており，大規模データにも対応できる機能を備えていることから，利用している．従って，ここではCDL.nl記述と並行してRDF形式でもデータを保存，管理する．

意味的類似性に基づく検索は，上記のような単語ベースの条件緩和以外にも関係子についても可能であり，今後どのような順序の条件緩和戦略が有効かを確認する必要がある．更に次の段階では，「著者は本を書いた人」といった一般知識があれば，「内田裕士は"UNL"の本を書いた．」というテキストとCDL.nl記述があれば，「"UNL"の本の著者は誰か？」という問い合わせに対し，"内田裕士"と答えることが出来るようになると期待できる．

7．むすび

Web3.0の名前は聞かれるようになってきたが，その具体的内容はまだ定まっている訳ではない．しかし，意味的検索や意味計算はその重要な機能になると考えられている．本稿はこのような次世代Webの意味計算の基盤となるCDL／CWLについて紹介した．Webは情報流通，共有，共創の重要なプラットフォームになっており，今後ともその役割の重要性は増していく．日本でもWeb関連の技術開発は数多く行われているものの，Web基盤構築に貢献するような研究開発は皆無と言える状態である(注3)．今後のWebの重要性を考えると，これは日本の情報技術の国際的位置にとって由々しき事態である．ここでのCDL／CWLは日本発の技術であり，次世代Web基盤構築に貢献しようとするものである．

W3Cのインキュベータグループで国際標準化に向けた議論を行っているのだが，実は英語を主とする国の関心は薄いのが実情である．これらの国では英語で事足りるので，多国語対応のコンピュータ・エスペラント語とも言えるCDL／CWLへの必要性を認識しにくいようである．(実はそうではなく，テキストの表す意味概念を正しくコンピュータに伝える媒介記述言語として必要である．)非英語国の日本がリードして，広く世界に貢献すべき課題であると思う．また，少子化により人口減少が進む日本で創造力を保つために，安易に移民に頼るといった解決策を採ることは難しいと思う．そうであるなら，言語の壁を超えて情報の流通，共有を可能にする情報技術によって，世界の智を活用できる環境を整備することは，日本の今後の重要な戦略的方策になると思う．

謝辞：

CDL／CWLの開発・国際標準化活動はNPO法人セマンティックコンピューティング研究開発機構(ISeC)を中心に行っている．この運営を行っている安原宏氏に感謝します．またUNDL財団で内田と共にUNLの開発を担ってきた朱美英さんに感謝します．

(注3) 日本からのW3Cでの標準化へのこれまでの貢献は，携帯Webブラウザに関する技術(NTTドコモ中心)，XML(日本だけではないが，日本IBM等が寄与)くらいに留まっている．

参 考 文 献

[CWL 08] http://www.w3.org/2005/Incubator/cwl/XGR-cwl/

[CDL.nl 06] Specification of CDL.nl, ISeC 資料 (2006), http://www.instsec.org/tr/

[石塚06] 石塚満：自然言語テキストの共通的概念記述，人工知能学会誌，Vol.21，No.6，pp.691-698（2006.11）
[Palmer 05] M. Palmer, D. Gildea and P. Kingsbury : The Proposition Bank : An Annotated Corpus of Semantic Roles, Computational Linguistics, Vol.31, No.1, pp.71-106（2005）
[SPARQL] http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/
[Uchida 90] Hiroshi Uchida and Meiying Zhu : Interlingua, International Symposium on Multilingual Machine Translation '90（MMT'90），(1990)
[Uchida 05] H. Uchida, M. Zhu and T. G. D. Senta : UNL-Universal Networking Language, UNDL Foundation (2005)
[UNDL] http://www.undl.org/
[Yan 08] Yulan Yan, Yutaka Matsuo, Mitsuru Ishizuka and Toshio Yokoi : Annotating an Extension Layer of Semantic Structure for Natural Language Text, Proc. 2nd IEEE Int'l Conf. on Semantic Computing, pp.174-181, Santa Clara, CA, USA,（2008.8）
[Yokoi 05] T. Yokoi, H. Yasuhara, H. Uchida, M. Zhu and K. Hashida : CDL（Concept Description Language）: A Common Language for Semantic Computing, Online Proc. WWW2005 Workshop on the Semantic Computing Initiative（SeC2005），Makuhari, Japan（2005.5）
[横井06] 横井俊夫：セマンティックコンピューティング－知的システム・知的環境の設計原理，人工知能学会誌，Vol.21，No.6，pp.683-690（2006.11）

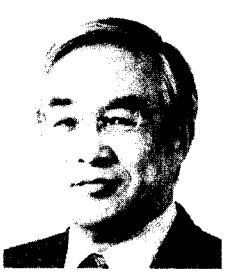
（2009年7月10日　受付）

[問い合わせ先]
〒113-8656　東京都文京区本郷7-3-1
東京大学 情報理工学系研究科
石塚　満
TEL：03-5841-6347
FAX：03-5841-8570
E-mail：ishizuka@i.u-tokyo.ac.jp

著者紹介

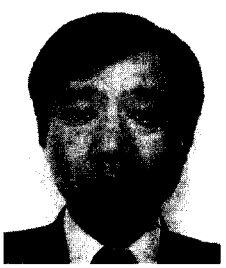
石塚　満（いしづか　みつる）［非会員］

1971年東京大学工学部電子工学科卒業．1976年同大学院工学系研究科博士修了．工学博士．同年NTT入社，横須賀研究所勤務．1978年東京大学生産技術研究所・助教授．（1980-81年Purdue大学客員准教授）．1992年同大学工学部電子情報工学科・教授．2001年大学院情報理工学系研究科電子情報学専攻・教授．2005年同研究科創造情報学専攻（電子情報学専攻兼任）．研究分野は人工知能，Webインテリジェンス，意味計算，生命的エージェントによるマルチモーダルメディア．IEEE，AAAI，人工知能学会（元会長），電子情報通信学会，情報処理学会，映像情報メディア学会，画像電子学会の各会員．



横井　俊夫（よこい　としお）［非会員］

1965年東京大学工学部卒業．1966年電気試験所（現在：産業技術総合研究所）．1982年第五世代コンピュータプロジェクトの推進に従事．1987年電子化辞書プロジェクトの推進，運営に従事．1995年フィリピン国にてODAプロジェクトの推進，指導に従事．1997年東京工科大学工学部教授．1999年同大学メディア学部教授．2008年（財）日本特許情報機構特許情報研究所顧問．工学博士．



内田　裕士（うちだ　ひろし）［非会員］

1970年大阪大学理学部物理学科卒業．1971年富士通入社．1986年日本電子化辞書研究所兼務．1988年国際情報化協力センター機械翻訳研究所兼務．1996年国連大学高等研究所入所．2001年UNDLファウンデーション理事．研究分野はコンパイラ，ソフトウェアエンジニアリング，人工知能，自然言語処理．